Content-Variant Reference Image Quality Assessment via Knowledge Distillation

Guanghao Yin[1]*, Wei Wang[2]†, Zehuan Yuan[2], Chuchu Han[3], Wei Ji[4], Shouqian Sun[1], Changhu Wang[2],

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China,
[2] Bytedance Inc, China, [3] Huazhong University of Science and Technology, Wuhan, China,
[4]National University of Singapore, Singapore
{ygh_zju, ssq}@zju.edu.cn, {wangwei.frank, yuanzehuan,changhu.wang}@bytedance.com, hcc@hust.edu.cn, jiwei@nus.edu.sg

Abstract

Generally, humans are more skilled at perceiving differences between high-quality (HQ) and low-quality (LQ) images than directly judging the quality of a single LQ image. This situation also applies to image quality assessment (IQA). Although recent no-reference (NR-IQA) methods have made great progress to predict image quality free from the reference image, they still have the potential to achieve better performance since HQ image information is not fully exploited. In contrast, full-reference (FR-IQA) methods tend to provide more reliable quality evaluation, but its practicability is affected by the requirement for pixel-level aligned reference images. To address this, we firstly propose the content-variant reference method via knowledge distillation (CVRKD-IQA). Specifically, we use non-aligned reference (NAR) images to introduce various prior distributions of high-quality images. The comparisons of distribution differences between HQ and LQ images can help our model better assess the image quality. Further, the knowledge distillation transfers more HQ-LQ distribution difference information from the FR-teacher to the NAR-student and stabilizing CVRKD-IQA performance. Moreover, to fully mine the local-global combined information, while achieving faster inference speed, our model directly processes multiple image patches from the input with the MLP-mixer. Cross-dataset experiments verify that our model can outperform all NAR/NR-IQA SOTAs, even reach comparable performance with FR-IQA methods on some occasions. Since the content-variant and non-aligned reference HQ images are easy to obtain, our model can support more IQA applications with its relative robustness to content variations. Our code and more detail elaborations of supplement are available: https://github.com/guanghaoyin/CVRKD-IQA.

Introduction

The target of objective image quality assessment (IQA) is to quantify the visual distortion and produce the perceptive quality score of the image. The accurate IQA method is quite important to guide many downstream tasks of image processing, such as image restoration (Banham and Katsaggelos 1997), super-resolution (Dong et al. 2015), etc.

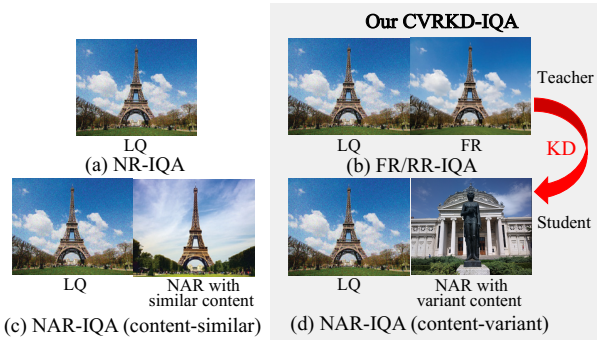


Figure 1: Based on the way of using HQ reference images, previous IQA methods are divided into (a) NR-IQA, (b) FR/RR-IQA, (c) NAR-IQA (content-similar). We propose the CVRKD-IQA method, which firstly uses the knowledge distillation to transfer HQ-LQ distribution difference information from (b) FR-IQA to (d) NAR-IQA (content-variant).

Recent studies on the human visual system (Sheikh and Bovik 2006; Ponomarenko et al. 2009) have shown that humans tend to compare images than directly judging an image. Provided with a high-quality (HQ) reference image, humans can make a more accurate and consistent evaluation about the quality of the distorted image (Ponomarenko et al. 2009). Based on the way of using HQ reference images, IQA methods are generally divided into three types: no-reference (NR) IQA, reduced-reference (RR) IQA, and full-reference (FR) IQA. Specifically, NR-IQA methods (Bosse et al. 2017; Su et al. 2020) (Fig. 1(a)) only use LQ images as input to directly measure image quality. FR/RR-IQA methods (Rehman and Wang 2012; Cheon et al. 2021) (Fig. 1(b)) utilize the complete or partial information of the pixel-aligned HQ reference images. Moreover, previous DCNN (Liang et al. 2016) uses the non-aligned image for IQA reference (NAR-IQA), which have similar contents but are not pixel-aligned with the LQ image (Fig. 1(c)).

Recently, there are several different attempts for NR-IQA methods to achieve promising performance, such as involv-

*This work was performed while Guanghao Yin worked as an intern at ByteDance.
†Equal Contribution.

ing a larger-scale database (Lin, Hosu, and Saupe 2019) or using a pretrained feature extractor (Su et al. 2020). Those NR-IQA methods still have the potential for better performance, because they focus more on mining the quality features of LQ image to better fit the labeled scores, but access little HQ image information. Humans can directly judge the quality of one image because they have learned various prior knowledge of HQ-LQ distribution differences before. Hence, we consider involving more explicit HQ prior distribution in the training and inference phases. As proved by previous works (Ponomarenko et al. 2009; Liang et al. 2016), the IQA scores predicted by reference-based IQA methods tend to be more consist with humans than those of NR-IQA methods. However, one strong requirement limits the application of previous FR/NAR-IQA models: their reference images must be pixel-wise aligned or have similar contents with LQ image, which are often unavailable in real scenarios. Thus, on the one hand, we attempt to loose this strong restriction and use content-variant HQ images for reference, since HQ images are available anywhere. On the other hand, inspired by the recent success of cross-modal knowledge distillation (KD) (Lan, Zhu, and Gong 2018; Porrello, Bergamini, and Calderara 2020), we consider to transfer more HQ-LQ difference information from FR-IQA model to NAR-IQA model via KD, which helps NAR-IQA model achieve more accurate and stable performance.

In this paper, we propose the first content-variant reference method via knowledge distillation (CVRKD-IQA) to assess image quality. The structure of our CVRKD-IQA is shown in Fig. 2. It consists of two parts: the FR-teacher and NAR-student. They have the same network structure while using different HQ reference images, *i.e.*, the pixel-aligned FR images and content-variant NAR images. For each network branch, the dual-path encoder separately extracts discriminative vectors from the LQ image itself and the HQ-LQ distribution difference. To transfer distribution difference knowledge from FR-teacher to NAR-student, we apply the offline knowledge distillation. The knowledge distillation can also constrain the NAR-student to focus more on useful HQ-LQ distribution difference representation by learning from FR-teacher, and reduce the impact of reference image changes to stabilize the NAR-IQA performance. Specifically, we first train the FR-teacher and fix its parameters, then the layers of the FR-teacher are employed to guide the training of NAR-student. Moreover, to effectively mine the global and local information of the image, our model directly processes a fixed number of image patches sampled from the full image with the classic MLP-mixer (Tolstikhin et al. 2021), which also keeps faster network inference. It should be noted that the FR-teacher is only for training and the NAR-student is applied for testing.

We have conducted extensive comparisons between our model and FR/NR/NAR-IQA SOTAs. Experimental results show that not only our FR-teacher can produce accurate IQA scores, but also our NAR-student can significantly outperform existing NR/NAR-IQA methods, especially on the large-scale real IQA dataset. On some occasions, our NAR-student can reach comparable performance with some common FR-IQA methods, such as PSNR and LPIPS. It further demonstrates that the proposed strategy transfers HQ-LQ difference prior knowledge from FR-teacher to NAR-student. Moreover, when using different content-variant HQ images, our NAR-student can still keep the relatively stable performance, which proves the robustness of our method.

In summary, our overall contribution is summarized as:

- We propose the first content-variant reference method via knowledge distillation (CVRKD-IQA), which introduces more HQ-LQ distribution difference knowledge.
- With the guidance of non-aligned reference image and knowledge distillation, our model significantly outperforms existing NR/NAR-IQA methods on synthetic and authentic IQA datasets. On some occasions, our model even reaches comparable results with FR-IQA metrics.
- Our model can directly use content-variant HQ images for reference, which can loose the restrictions of previous pixel-aligned or content-similar reference images.

Related Work

Image Quality Assessment. The target of objective IQA is to accurately acquire the consistent quality of one image with human views. According to the involvement of reference images, the objective IQA can be generally classified into three types: full-reference (FR), reduced-reference (RR), and no-reference (NR) IQA methods.

In general, FR/RR-IQA simulates the sensitivity of the human visual system to different image signals (Sheikh and Bovik 2006), including information-theoretic criterion (Wang et al. 2004), structural information (Zhang et al. 2011), etc. The FR-IQA methods perform their quality measurements based on point-by-point comparisons between pixels. And FR-IQA has been widely applied as the perceptual metric for downstream tasks of image proceeding. The most commonly and widely used FR metrics are the PSNR and SSIM (Wang et al. 2004), which are convenient for optimization. Recently, learning-based FR-IQA methods (Prashnani et al. 2018; Ding et al. 2021) have achieved significant improvement. The most current IQT model (Cheon et al. 2021) involves the visual transformer with extra quality and position embeddings to achieve the best performance for the FR-IQA task. Different from FR-IQA, the RR-IQA method (Rehman and Wang 2012) utilizes only parts of the FR image information. Since RR-IQA has the advantages of lower calculation expense and faster speed, it's commonly applied in the image transmission system.

For NR-IQA, CNN-based methods (Bosse et al. 2017; Wu et al. 2020; Su et al. 2020) have significantly outperformed handcrafted statistic-based approaches (Xu et al. 2016) by directly extracting discriminative features from LQ images. Due to distortion diversity and content changes, the recent trend of NR-IQA (Li, Jiang, and Jiang 2020) is to involve semantic prior information by using pretrained models on classification databases, *i.e.*, ImageNet (Deng et al. 2009). And Su *et al.* (Su et al. 2020) propose a dynamic hyper-network to adaptively adjust the quality prediction parameters based on image content. Recently, You *et al.* (You and Korhonen 2021) introduce the visual transformer for the NR-IQA task.

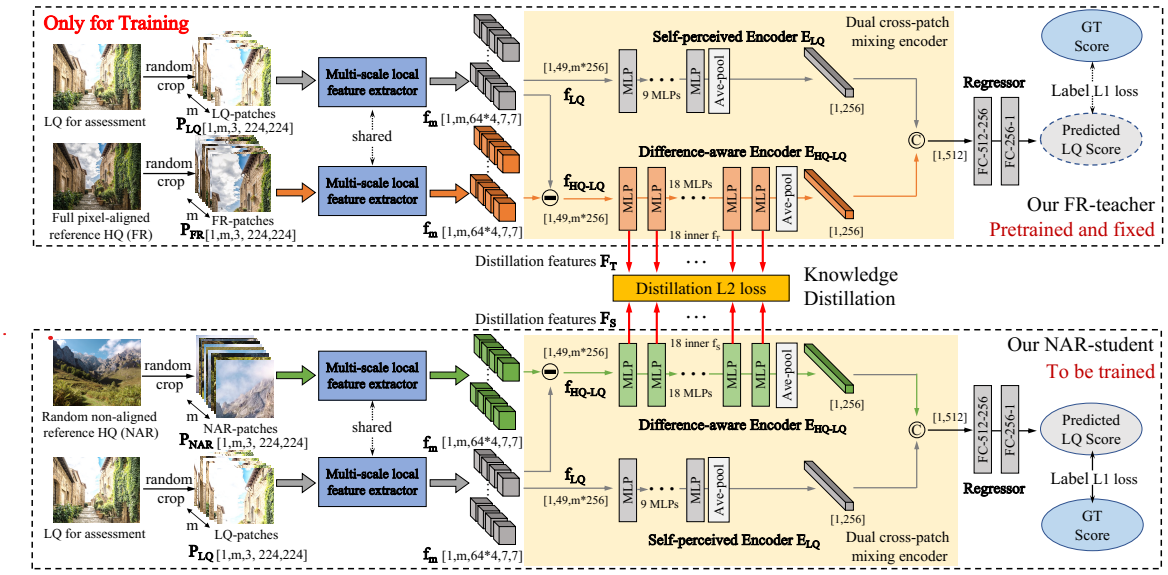


Figure 2: Model overview of our CVRKD-IQA. It consists of FR-teacher and NAR-student with the same structure. For better conducting local-global quality description, we use the multi-patches randomly cropped from LQ and reference images as inputs. Note that the FR-teacher is pretrained and fixed only for distillation and the trained NAR-student is applied for testing.

However, FR-IQA methods tend to provide more reliable quality evaluation than NR-IQA models (Zhang et al. 2011).

Since pixel-aligned FR images are not always available, DCNN (Liang et al. 2016) defines a new task named Non-aligned Reference IQA (NAR-IQA), which uses a reference image with similar scene but is not well aligned with the LQ image. Nevertheless, the images with similar scenes are still not always easy to get. Recently, Ma *et al.* (Ma et al. 2017) form the quality-discriminable image pairs to help rank the IQA scores, and Guo *et al.* (Guo et al. 2021b) introduces the pseudo images for reference. However, those methods still need to manually form their reference images. In this paper, we attempt to use content-variant HQ images for reference.

Knowledge Transfer via Distillation. Transferring knowledge from one model to another has been a long line of research. Ba and Caruana (Ba and Caruana 2014) successfully increase the accuracy of a shallow neural network by training it to mimic a deeper one and penalize the difference of logits between them. Hinton et al. (Hinton, Vinyals, and Dean 2015) revive this idea under the name of knowledge distillation (KD) that trains a student model to match the distribution of a teacher model. Although the KD strategy was primarily proposed for model compression (Lan, Zhu, and Gong 2018), many recent works have extended the cross-modal distillation to multi-modal visual tasks, such as action recognition (Garcia, Morerio, and Murino 2018), person re-identification (Porrello, Bergamini, and Calderara 2020) or depth estimation (Gupta, Hoffman, and Malik 2016), where the knowledge of different modals are transferred between different network branches. In this paper, we make the first attempt to transfer more HQ-LQ difference prior information from the FR-IQA to the NAR-IQA via KD. Experiments prove that distillation operation can further help our NAR-student achieve more accurate and stable performance.

Proposed Method

In this section, we will introduce the structure of CVRKD-IQA and explain how to transfer the distribution difference knowledge from FR-teacher to NAR-student.

Network Architecture

Overall Architecture. As shown in Fig. 2, our model consists of two parts: FR-teacher N_T and NAR-student N_S. Both of them use two types of images: the distorted LQ image I_{LQ} for assessment and the HQ image I_{HQ} for reference. The FR-teacher N_T and NAR-student N_S have the same structure. The only difference between them is I_{HQ}, where FR-teacher N_T uses pixel-aligned I_{FR} and NAR-student N_S uses random non-aligned content-variant I_{NAR}.

Image quality is perceived by both local degradation and global information. Recently, some representative IQA methods (Su et al. 2020; Cheon et al. 2021) usually use one local image patch as input, and average or reweight the predicted scores of each local patch to get final results. However, this operation does not make full and effective use of local-global combined information. Compared with single image patch input, multiple patches input can more effectively provide information on both local fine-grained distortion from the single patch and global coarse-grained composition cross patches at one time. Hence, our model uses 2 sets of m multi-patches as inputs, *i.e.*, $P_{LQ} = \{p_{LQ_i}\}(i = 1, ...m)$ and $P_{HQ} = \{p_{HQ_i}\}(i = 1, ...m)$, which are randomly cropped from I_{LQ} and I_{HQ}. It should be noted that $\{P_{LQ}, P_{FR}\}$ are still pixel-aligned for the FR-teacher.

To combined the advantages of NR-IQA and FR-IQA methods, our model attempts to mine the local-global combined features from the LQ image itself and HQ-LQ distribution difference. Moreover, the multi-scale feature extrac-

tion should also be conducted to better describe the local distortion. To achieve those, we design three modules for our model: (1) the multi-scale local feature extractor; (2) dual cross-patch mixing encoders; (3) a full-connected regressor;

Multi-scale Local Feature Extractor. First, following (Li, Jiang, and Jiang 2020; Cheon et al. 2021), the pretrained CNN backbone on the image classification task is applied as the perceptual feature extractor. Thus, we use the pretrained ResNet50 (He et al. 2016) on ImageNet (Deng et al. 2009) to process input patches $\{P_{LQ}, P_{HQ}\}$. Since features from different scale layers are important to capture local distortions (Su et al. 2020; Guo et al. 2021a), we design a multi-scale feature extractor. Specifically, four scale features from conv2_9, conv3_12, conv4_18, conv5_9 layers of ResNet50 are processed by 1×1 convolution and global average pooling. Those four feature maps with the same size ($[m, 64, 7, 7]$) are concatenated as f_m in channel-wise ($[m, 256, 7, 7]$) to describe local distortions.

Dual Cross-patch Mixing Encoder. Then, we use cross-patch mixing encoders to extract the self-perceived feature f_{LQ} and the HQ-LQ difference-aware features f_{HQ-LQ}, respectively. To effectively explore local-global combined information from multi-patches input, we build our encoders with the classic MLP-mixer (Tolstikhin et al. 2021), which has a simpler architecture and faster speed than the visual transformer (Vaswani et al. 2017). Different from the original MLP-mixer (Tolstikhin et al. 2021) for classification tasks fed with spatial image tokens, our encoders operate on the multi-scale features f_m extracted from multi-patches input. Each MLP module consists of two blocks: the first one is patch-mixing MLP block, which exchanges inner information between transposed local features of multi-patches; the next one is channel-mixing MLP block, which allows global information communication between multi-patches and multi-scales. Since mining the distribution difference between I_{LQ} and I_{HQ} is much more difficult than I_{LQ} perceive feature extraction, we design deeper encoder E_{HQ-LQ} with 18 stacked MLP modules and the encoder E_{LQ} uses only 9 stacked MLP modules. The final layer normalization and global average pooling convert feature maps to vectors. Two cross-patch vectors ([256, 1]) from the dual-path encoder are concatenated as ([512, 1]) for quality regression prediction.

Regressor for Quality Prediction. Since the regressor is simply mapping the output vectors of the dual-path encoder to labeled quality scores, we design a small network for faster quality prediction. The regressor consists of two fully-connected layers with 512-256, 256-1 channels to predict the final quality score of the input LQ image.

Knowledge Distillation from FR-IQA to NAR-IQA

Considering that our goal is to transfer more HQ-LQ distribution knowledge, and better constrain NAR-student for useful HQ-LQ distribution difference representation, we perform the distillation operation between the difference-aware encoders E_{HQ-LQ} of FR-teacher and NAR-student.

To obtain a well-performed FR-teacher, we do not jointly train FR-teacher and NAR-student, but apply an offline distillation scheme. First, we randomly crop two multi-patch sets $\{P_{LQ}, P_{FR}\}$ from the LQ-FR image pair $\{I_{LQ}, I_{FR}\}$ as the input. The FR-teacher $N_T(\cdot;\theta_1)$ is optimized by L_1 loss between predicted score $\hat{y}_t$ and ground-truth y as:

$$L_{T_l} = \frac{1}{N}\sum_{i=1}^{N} ||y_i - N_T(P_{LQ}^{(i)}, P_{FR}^{(i)}; \theta_1)||_1. \quad (1)$$

Then, we fix the parameters of the trained FR-teacher. The NAR-student is supervised by the guide of FR-teacher and human labeled scores in the second step of training. Except the paired $\{P_{LQ}, P_{FR}\}$ for FR-teacher input, NAR-student should also be fed with the non-aligned $\{P_{LQ}, P_{NAR}\}$, where P_{NAR} consists of m randomly cropped patches from another non-aligned reference HQ image. We attempt to transfer more prior knowledge of HQ-LQ distribution difference from FR-teacher to NAR-student. Hence, all 18 inner features $F_T = \{f_{T_j}\}(j = 1, 2, ...18)$ of the difference-aware encoder of FR-teacher are applied to guide the training of NAR-student. The L_2 loss is used as the distillation loss L_{S_d} to transfer knowledge to corresponding layer features $F_S = \{f_{S_j}\}(j = 1, 2, ...18)$ of NAR-student:

$$L_{S_d} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K=18} ||f_{T_j}^{(i)} - f_{S_j}^{(i)}||_2. \quad (2)$$

Except distillation loss, the label loss L_{S_l} between predicted results $\hat{y}_s$ and labeled ground-truth y is also applied to optimize the NAR-student $N_S(\cdot;\theta_2)$:

$$L_{S_l} = \frac{1}{N}\sum_{i=1}^{N} ||y_i - N_S(P_{LQ}^{(i)}, P_{NAR}^{(i)}; \theta_2)||_1. \quad (3)$$

And the final loss L_S for NAR-student is combined by the distillation loss L_{S_d} in Eq. 2 and label loss Eq. 3 as:

$$L_S = L_{S_d} + L_{S_l}. \quad (4)$$

With the guidance of knowledge distillation, our NAR-student effectively learns more HQ-LQ difference knowledge and keeps the stability with different NAR images. In real scenarios, when the pixel-aligned FR image is unavailable but HQ images are easy to get, our NAR-student can directly use any non-aligned HQ image for reference.

Experiments

Experimental setting

In this paper, all comparisons of FR/NR/NAR-IQA methods and ablation studies follow this setting.

Datasets. For IQA training datasets, we follow (Cheon et al. 2021) to choose the commonly used synthetic Kaddid-10K (Lin, Hosu, and Saupe 2019), which contains 10125 LQ-FR pairs. The cross-dataset evaluations are conducted on 3 synthetic datasets, *i.e.*, LIVE (Sheikh, Sabir, and Bovik 2006), CSIQ (Larson and Chandler 2010), TID2013 (Ponomarenko et al. 2015), which separately contains 779, 886 and 3000 LQ-FR pairs with traditional distortions. Moreover, we also evaluate on large-scale authentic KonIQ-10K dataset (Hosu et al. 2020), containing 10073 real-distorted LQ images without FR images. Except for IQA datasets, our

IQA Type	Method	LIVE			CSIQ			TID2013			KonIQ-10K		
		SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
FR-IQA	PSNR	0.873	0.865	0.680	0.810	0.819	0.601	0.687	0.677	0.496	-	-	-
	MAD (Larson and Chandler 2010)	0.967	0.968	0.842	0.947	0.950	0.797	0.781	0.827	0.604	-	-	-
	WaDIQaM-FR (Bosse et al. 2017)	0.947	0.940	0.791	0.909	0.901	0.732	0.831	0.834	0.631	-	-	-
	PieAPP (Prashnani et al. 2018)	0.919	0.908	0.750	0.892	0.877	0.715	0.876	0.859	0.683	-	-	-
	LPIPS (Zhang et al. 2018)	0.932	0.934	0.765	0.876	0.896	0.689	0.670	0.749	0.497	-	-	-
	DISTS (Ding et al. 2021)	0.954	0.954	0.811	0.929	0.928	0.767	0.830	0.855	0.639	-	-	-
	IQT (Cheon et al. 2021)	0.970	-	0.849	0.943	-	0.799	**0.899**	-	**0.717**	-	-	-
	Our FR-teacher	**0.973**	**0.969**	**0.853**	**0.964**	**0.964**	**0.829**	0.890	**0.886**	0.698	-	-	-
NR-IQA	CNNIQA (Kang et al. 2014)	0.653	0.656	0.485	0.649	0.660	0.482	0.476	0.404	0.283	0.278	0.285	0.183
	WaDIQaM-NR (Bosse et al. 2017)	0.855	0.855	0.656	0.716	0.750	0.527	0.585	0.610	0.416	0.382	0.386	0.261
	HyperIQA (Su et al. 2020)	0.908	0.903	0.730	0.802	0.858	0.611	0.686	0.721	0.490	0.332	0.338	0.233
	TRIQ (You and Korhonen 2021)	0.909	0.910	0.729	0.807	0.862	0.615	0.684	0.731	0.500	0.371	0.371	0.259
	LinearityIQA (Li, Jiang, and Jiang 2020)	0.910	0.906	0.738	0.815	0.873	0.629	0.688	0.694	0.491	0.361	0.361	0.254
NAR-IQA	DCNN (Liang et al. 2016)	0.752	0.756	0.594	0.721	0.716	0.583	0.473	0.492	0.346	0.258	0.256	0.147
	WaDIQaM (Bosse et al. 2017)-NAR w/ KD	0.897	0.894	0.707	0.799	0.851	0.613	0.670	0.694	0.493	0.362	0.364	0.258
	IQT (Cheon et al. 2021)-NAR w/ KD	0.908	0.906	0.728	0.802	0.860	0.624	0.680	0.707	0.499	0.372	0.372	0.269
	Our NAR-student	**0.913**	**0.917**	**0.748**	**0.829**	**0.872**	**0.655**	**0.691**	**0.733**	**0.501**	**0.416**	**0.413**	**0.287**

Table 1: Model comparisons on synthetic LIVE, CSIQ, TID2013, and authentic KonIQ-10K when training on synthetic Kaddid-10K. We also extend two FR-IQA methods (WaDIQaM, IQT) to NAR-IQA via knowledge distillation (KD). It's clear that our NAR-student can outperform all NR/NAR-IQA methods, especially on the large-scale authentic KonIQ-10K with real unknown distortions. On TID2013, our NAR-student reaches comparable and even better performance than PSNR and LPIPS.

NAR-student still need non-aligned HQ reference images. The 900 training and 100 testing HQ images of DIV2K_HR dataset (Agustsson and Timofte 2017) are randomly sampled at the training and testing stages of NAR-student.

Evaluation Criterias. The Spearman's rank order correlation coefficient (SRCC), Pearson's linear correlation coefficient (PLCC) and Kendall rank order correlation coefficient (KRCC) are employed to measure prediction monotonicity and prediction accuracy. The higher value indicates better performance. For PLCC, the logistic regression correction is also applied according to (Antkowiak et al. 2000).

Implementation Details. Data augmentation including horizontal flip and random rotation is applied during the training. All patches are randomly cropped from the RGB image. The batch size b is set as 32. The input patch number m is set as 10 and the patch size is set as $224 \times 224 \times 3$ to cover more local-global combined information. The number k of distilled layers in the encoder E_{HQ-LQ} is set to 18. Moreover, the initial learning rate α is 2×10^{-5} and the ADAM optimizer with weight decay 5×10^{-4} is applied. All the experiments were conducted on NVIDIA Tesla-V100 GPUs.

Comparisons with the State-of-the-art Methods

Here, we will present the accuracy and generalization comparisons between our model and existing FR/NR/NAR-IQA methods. Specifically, our FR-teacher and NAR-student are separately compared with FR-IQA SOTAs *i.e.*, (Ding et al. 2021; Cheon et al. 2021) and recent best performed NAR/NR-IQA SOTAs *i.e.*, (Liang et al. 2016; Li, Jiang, and Jiang 2020). Moreover, we also extend two FR-IQA methods (WaDIQaM (Bosse et al. 2017), IQT (Cheon et al. 2021)) to NAR-IQA via knowledge distillation. Specifically, we use the pretrained WaDIQaM and IQT as teachers under FR-IQA settings. And we obtain the corresponding student models by changing the reference input from FR images to NAR images. Following our strategy, the knowledge distillation is also applied in those models and we get the WaDIQaM-NAR and IQT-NAR w/ KD. Since we follow the commonly used experimental settings of FR-IQA methods (Ding et al. 2021; Cheon et al. 2021), we directly use those published FR-IQA results. For fair comparisons between FR/NR/NAR methods, we retrain those NR/NAR-IQA SOTAs with the same experimental setting as ours.

The results of the four datasets are shown in Table 1. For the FR-IQA setting, since the authentic Kaddid-10K doesn't provide FR images, the FR-IQA comparisons are only conducted in 3 synthetic datasets. It can be seen that our FR-teacher outperforms all FR-IQA methods on LIVE and CSIQ, and it is also ranked in the top two in all benchmarks with the marginal gap on larger TID2013. Hence, the trained FR-teacher is good enough as the distillation teacher. When pixel-aligned FR images are not provided, our NAR-student outperforms existing NR/NAR-IQA models on all 4 testsets, which shows that the proposed strategy can improve the IQA performance. What's more, the comparisons about distilled WaDIQaM-NAR and IQT-NAR further prove this point. It should be noted that although trained on the synthetic Kaddid-10K, our NAR-student achieves significant improvement than NR/NAR-IQA SOTAs on the authentic KonIQ-10K. Moreover, on the synthetic TID2013, our NAR-student reaches comparable and even better performance than the commonly used FR-IQA methods, such as PSNR and LPIPS (Zhang et al. 2018).

Runtime vs. Performance

To compare the efficiency of our NAR-IQA model with other NR/NAR methods in the inference stage, we report the average runtime of IQA for a distorted image with the num-

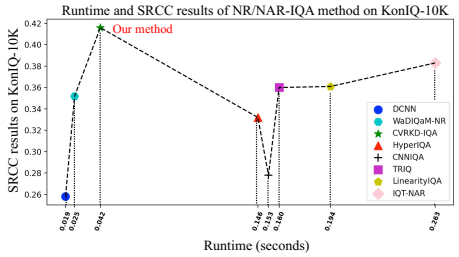


Figure 3: Runtime vs. performance of NR/NAR-IQAs on the real-distorted KonIQ-10k dataset with the Tesla-V100 GPU.

ber of patches $m = 10$ in Fig. 3. On real-distorted KonIQ-10K, our CVRKD-IQA significantly outperforms NR/NAR-IQA SOTAs and satisfies the real-time requirement (about 24 images per second), while transformer-based TRIQ (You and Korhonen 2021) and IQT-NAR (Cheon et al. 2021) cost much more inference time. All experiments were conducted on NVIDIA Tesla-V100 GPU.

Ablation Study

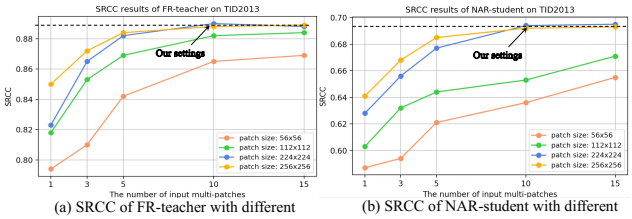
Effect of Knowledge Distillation (KD) and Non-aligned Reference Images (NAR). First, we separately analyze the effects of knowledge distillation (KD) and non-aligned reference (NAR) images in Table 2. It should be noted that NAR is the pre-condition of KD. If the NAR image is not provided, the NAR-student cannot mine the HQ-LQ distribution difference, thus cannot learn the transferred knowledge from the FR-teacher. Hence, we evaluate 3 types of KD and NAR configurations. Except SRCC metrics, we also present the standard deviations (Std) of 10 SRCC results tested with different misaligned reference images. From results in Table 2, we can make the following analyses:

- We first remove the difference-aware encoder to train NR-student baseline under NR-IQA setting without NAR or KD. It's clear the NR-student baseline achieves worse performance, especially in real-distorted KonIQ-10K.
- When NAR images are available, the NAR-student w/o KD benefits from the HQ-LQ distribution difference to outperform the NR-student baseline. However, the performance of NAR-student w/o KD is the most unstable with the highest Std values. It means various contents of different NAR images increase the training difficulty.
- When provided with more HQ-LQ difference knowledge from the FR-teacher by KD, our final NAR-student achieves the best performance, especially the 33% SRCC improvements than NR-student baseline in KonIQ-10K. Moreover, Std results of our final NAR-student decreased to 0.004, which proves the great importance of KD to stabilize the NAR-IQA performance.

Effectiveness of Multi-patches. To make full and effective use of local-global combined information, our method directly processes multi-patches and fuses the cross-patch features via the MLP-mixer. As shown in Fig. 4, we gradually increase the patch number (1, 3, 5, 10, 15) and the patch size (56, 112, 224, 256) to analyze the effects of multi-patches. It's clear that both FR-teacher and NAR-student

Model	Configs NAR	KD	TID2013 SRCC ↑ ± Std ↓	KonIQ-10K SRCC ↑ ± Std ↓
NR-student baseline	×	×	0.631 ± **0.002**	0.317 ± **0.003**
NAR-student w/o KD	✓	×	0.679 ± 0.056	0.352 ± 0.072
NAR-student w/ KD	✓	✓	**0.691** ± 0.003	**0.416** ± 0.004

Table 2: SRCCs and the standard deviations (Std) of our student with different configurations of knowledge distillation (KD) and non-aligned reference (NAR) images.



(a) SRCC of FR-teacher with different patch number and size on TID2013

(b) SRCC of NAR-student with different patch number and size on TID2013

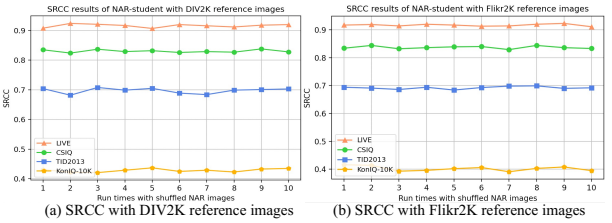
Figure 4: SRCC results with different multi-patch numbers and sizes on TID2013. It's clear that larger patch number and size can capture more local-global information.

benefit from larger patch number and size, because they can capture more local-global information to better describe the full-image quality. Considering the trade-off between inference efficiency and performance, the patch number is set to 10 and the patch size is set to 224×224.

Stability about Non-aligned Reference HQ Images. To further demonstrate the stability of our model when using content-variant HQ images for reference, we evaluate our NAR-student with more various HQ images. Specifically, not only DIV2K, we involve 2650 HQ images of Flikr2K (Timofte et al. 2017) as another non-aligned reference dataset. Since we randomly sample the HQ reference image for each LQ assessment, HQ images of each round are shuffled. Therefore, we also present results across 10 times. As shown in Fig. 5, we can make the following analyses:

- As shown in Fig. 5(a)(b), our NAR-student achieves relatively stable performance when using shuffled NAR images across 10 times on both DIV2K and Flikr2K.
- As shown in the comparisons between Fig. 5(a) and (b), our NAR-student also produces relatively similar SRCC results between DIV2K and Flikr2K.
- Those observations demonstrate that our NAR-student is stable and robust to content-variant NAR images.

Evaluation on Reference Image with Different Content. Now, there are 3 types of reference images: the pixel-aligned FR image, the NAR image with similar content and the random content-variant NAR image. How can we choose them properly based on content? Hence, we evaluate the distilled NAR-student with different types of HQ reference contents. Note that we follow (Liang et al. 2016) to synthesize the content-similar NAR image by applying affine transform to FR images (random scaling factor s and rotation θ from $[0.95, 1.05]$ and $[-5°, 5°]$). From results in Table 3 and



(a) SRCC with DIV2K reference images (b) SRCC with Flikr2K reference images

Figure 5: Stability evaluation of our NAR-student. (a)(b) show SRCC results of 10 times tests using randomly shuffled NAR images from DIV2K and Flikr2K, respectively.

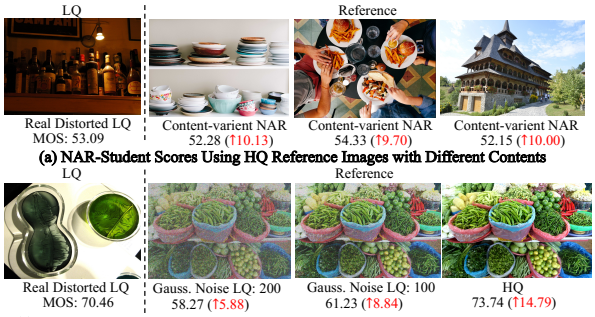


Figure 6: **Real-data examples** on KonIQ-10K, where the NAR-student uses reference images with (a) different contents and (b) different qualities. All scores are rescaled in [0, 100]. The red numbers are the decrease of the MOS error compared to the no-reference baseline.

Fig. 6(a), we can make the following analyses:

- As shown in the first two lines of Table 3, although the NAR-student is trained with NAR settings, it reaches comparable results with FR-teacher when using pixel-aligned FR images. This proves the NAR-student has learned transferred knowledge from the FR-teacher.
- As shown in the last two lines of Table 3, the performance of content-variant NAR images is slightly lower than content-similar NAR images. Since the stable and promising performance of our method has been proved in Fig. 6(a) and Fig. 5, we can use random HQ images for reference when content-similar images are unavailable.
- In real scenarios, we should choose the HQ reference image with aligned content as much as possible.

Evaluation on Reference Image with Different Quality. Although the HQ images are easy to obtain, we should still evaluate our NAR-student on NAR images with different qualities. Specifically, we first use various distorted reference images of synthetic TID2013 and authentic KonIQ-10K. Moreover, we choose 3 typical distortions to generate distorted reference images from DIV2K_HQ, *e.g.*, $\times 2$ downsample, random Gaussian noise with levels: [0,10], random JPEG compression with qualities: [0,10]. From results in Table 4 and Fig. 6(b), we can make the following analyses:

- As shown in Table 4 and Fig. 6(b), using reference images with higher quality can produce better results.

Model + Input Reference Image	LIVE SRCC	CSIQ SRCC	TID2013 SRCC
FR-teacher + Pixel-aligned FR	0.973	0.964	0.890
NAR-student + Pixel-aligned FR	0.958	0.937	0.846
NAR-student + Content-similar NAR	0.931	0.862	0.720
NAR-student + Content-variant NAR	0.913	0.829	0.691

Table 3: SRCC results using HQ reference images with different contents. For clear comparisons, we add the FR-teacher with pixel-aligned FR image as the upper-bound. It's clear that more aligned HQ images produce better results.

Reference	Distortion Type	TID2013 SRCC	TID2013 PLCC	KonIQ-10K SRCC	KonIQ-10K PLCC
KonIQ-10K	Authentic Distortions	0.671	0.711	0.392	0.393
TID2013	Synthetic Distortions	0.683	0.722	0.394	0.395
DIV2K	Gauss. Noise: [0, 10]	0.665	0.704	0.392	0.390
	JPEG Level: [0, 10]	0.668	0.702	0.401	0.401
	Downsample: $\times 2$	0.687	0.721	0.408	0.407
DIV2K_HQ		**0.691**	**0.733**	**0.416**	**0.413**

Table 4: The SRCC and PLCC results of our NAR-student with different qualities NAR images. The NAR-student are fixed, and we only change the types of reference images. It's clear NAR images with higher quality produce better results.

- As shown in the first two examples of Fig. 6(b), using severely distorted LQ images for reference just brings marginal improvements than the no-reference baseline.
- In real scenarios, we should choose the NAR image with high-quality as much as possible. Since the HQ images are easy to get, our NAR-student can directly use random obtainable HQ images for reference.

Conclusion

In this paper, we investigate the image quality assessment (IQA) problem with non-aligned reference (NAR) images. We propose the first content-variant NAR-IQA method via knowledge distillation, namely CVRKD-IQA. Our model uses various NAR images to introduce prior distributions of HQ images. The knowledge distillation further transfers more HQ-LQ distribution difference knowledge from the FR-teacher to the NAR-student and stabilizes IQA performance. We also use the multiple patches input to fully and effectively mine the multi-scale and local-global combined features. Extensive experiments have demonstrated that our CVRKD-IQA significantly outperforms existing NR/NAR-IQA methods, even reaches comparable performance with commonly used FR-IQA metrics. Evaluations with different NAR images also prove the relative robustness of our model, which can support more IQA applications with randomly obtainable HQ images. Moreover, the reference images with higher-quality and more aligned content produce better results. In future work, we will further explore more novel NAR-IQA architecture and knowledge distillation strategy.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Antkowiak, J.; Jamal Baina, T.; Baroncini, F. V.; Chateau, N.; FranceTelecom, F.; Pessoa, A. C. F.; Stephanie Colonnese, F.; Contin, I. L.; Caviedes, J.; and Philips, F. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000.

Ba, J.; and Caruana, R. 2014. Do Deep Nets Really Need to be Deep? *Advances in Neural Information Processing Systems*, 27.

Banham, M. R.; and Katsaggelos, A. K. 1997. Digital image restoration. *IEEE signal processing magazine*, 14(2): 24–41.

Bosse, S.; Maniry, D.; Müller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1): 206–219.

Cheon, M.; Yoon, S.-J.; Kang, B.; and Lee, J. 2021. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 433–442.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2021. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4): 1258–1281.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307.

Garcia, N. C.; Morerio, P.; and Murino, V. 2018. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision*, 103–118.

Guo, H.; Bin, Y.; Hou, Y.; Zhang, Q.; and Luo, H. 2021a. Iqma network: Image quality multi-scale assessment network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 443–452.

Guo, J.; Wang, W.; Yang, W.; Liao, Q.; and Zhou, J. 2021b. Subjective Opinions Matter: Controllable Image Quality Assessment Using Pseudo Reference Images. *arXiv preprint arXiv:2105.02464*.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on image processing*, 29: 4041–4056.

Kang, L.; Ye, P.; Li, Y.; and Doermann, D. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1733–1740.

Lan, X.; Zhu, X.; and Gong, S. 2018. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems*, 7528–7538.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006.

Li, D.; Jiang, T.; and Jiang, M. 2020. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, 789–797.

Liang, Y.; Wang, J.; Wan, X.; Gong, Y.; and Zheng, N. 2016. Image quality assessment using similar scene as reference. In *European Conference on Computer Vision*, 3–18. Springer.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.

Ma, K.; Liu, W.; Liu, T.; Wang, Z.; and Tao, D. 2017. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on image processing*, 26(8): 3951–3964.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30: 57–77.

Ponomarenko, N.; Lukin, V.; Zelensky, A.; Egiazarian, K.; Carli, M.; and Battisti, F. 2009. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4): 30–45.

Porrello, A.; Bergamini, L.; and Calderara, S. 2020. Robust re-identification by multiple views knowledge distillation. In *European Conference on Computer Vision*, 93–110. Springer.

Prashnani, E.; Cai, H.; Mostofi, Y.; and Sen, P. 2018. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1808–1817.

Rehman, A.; and Wang, Z. 2012. Reduced-reference image quality assessment by structural similarity estimation. *IEEE Transactions on image processing*, 21(8): 3378–3389.

Sheikh, H. R.; and Bovik, A. C. 2006. Image information and visual quality. *IEEE Transactions on image processing*, 15(2): 430–444.

Sheikh, H. R.; Sabir, M. F.; and Bovik, A. C. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11): 3440–3451.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3667–3676.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125.

Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A. P.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, J.; Ma, J.; Liang, F.; Dong, W.; Shi, G.; and Lin, W. 2020. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on image processing*, 29: 7414–7426.

Xu, J.; Ye, P.; Li, Q.; Du, H.; Liu, Y.; and Doermann, D. 2016. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on image processing*, 25(9): 4444–4457.

You, J.; and Korhonen, J. 2021. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1389–1393. IEEE.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8): 2378–2386.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.